X-Sender: J. Patrick Gary
Date: Fri, 17 Sep 2004 13:16:37 -0400
To: distribution suppressed
From: J. Patrick Gary
Subject: Fwd: Initial thunderhead cluster 10-GigE network testing
Cc: William.E.Fink

Dear Colleagues,
The info below in Bill Fink’s 16 Sept 04 email provides an indication of both:
• The type of 10 Gbps inter-cluster connectivity we've been planning via GSFC's L-Net efforts, and
• The end-to-end user/application-level throughput performance that we're targeting.

Unfortunately for the non-expert, the info below in Bill’s email is condensely presented; so let me try to explain a little.

With respect to (wrt) the type of 10 Gbps inter-cluster connectivity
John Dorband (935) gave Bill Fink (933) access to 16 cpu's in "essentially two different" clusters (actually it was 32 cpu's in the same cluster; but given the way Bill used them it was essentially two different clusters).

Note also that John could and would have given us access to more cpu's per cluster if we had asked. But we only asked for 10 to 16 cpu's per cluster because we expected, and Bill subsequently proved in these (and other prior) tests, that he could reach the maximum theoretical data carrying capability of the 10-Gbps network we placed between the clusters using Bill's software-based "nuttcp" user/application -level data generation and throughput measuring tool running between just 10 to 16 cpu-pair's from the respective clusters.

In the respective clusters, each cpu has a Gigabit Ethernet (GE) network interface card (NIC) that we used to connect to a local 1-GE switch which in turn has a 10-GE "uplink" to the L-Net. Note that each cluster's basic interconnection fabric (which may be Myinet, Fast Ethernet, GE, Infiniband, etc) among the cpu's is different from the "extra" 1-GE NIC's used for the L-Net connection.

In Bill diagram below, the individual cpu/1-GE-NIC's are labeled ethn228, ethn229, etc. (In Bill's diagram note within this sequence some numbers are not there, because temporarily those cpu's are or were in repair). Also note that for this test the L-Net consisted of simply one other 10-GE switch (the Force10 E300 10-GE switch shown in the middle of Bill's diagram); but for realism that 10-GE switch was located elsewhere within building 28. And within a few months we've planned that the L-Net will include several 10-GE switches enabling the above described type of L-Net inter-connections among several different clusters in different buildings at GSFC, ARC, JPL, UCSB/SIO, UIC, etc.

Wrt the end-to-end user/application-level throughput performance that we're targeting

Bill's nuttcp program runs in a Unix/Linux computer as a normal user/application; but he's programmed it to be multi-functional. For network stress testing purposes, we use nuttcp's features which:

- Generate data records in memory (note no disk I/O; nuttcp has features/uses that involve disk I/O, but they typically limit network stress testing);
- Varies the data record's size, number generated, and/or wall clock time used to generate data records;
- Varies various network "tunable" parameters such as MTU and transport window sizes;
- Transmits/receives the above data records as normal TCP/IP packets (or not-normal TCP/IP packets if we're testing new communications protocols, etc); and
- Measures "distributively" and reports "centrally" a number of performance metrics, particularly including both the wall clock time to transmit the user's data with the packet/network's overhead and the cpu use percent of the transmitting and receiving cpu's.

More info about nuttcp is available at ftp://ftp.lcp.nrl.navy.mil/pub/nuttcp/.

In Bill's reported performance data below, the meaning of some of the notations used are:

- tx228-245 means transmit from the memory of cpu 228 to the memory of cpu 245
- rx228-245 means receive at the memory of cpu 228 transmitted from the memory of cpu 245
- %TX identifies the cpu use percent of the transmitting cpu to perform

the test; and

- %RX identifies the cpu use percent of the receiving cpu to perform the test.

In some tests below, a single unidirectional data flow "streams" between each cpu-pair. In other tests below, two streams flow between each cpu-pair, but the flows are in different directions, a.k.a. bi-directional.

Please let me or Bill know if you need more explanation of the data presented below. Thanks.

Pat

From: Bill Fink
Subject: Initial thunderhead cluster 10-GigE network testing
To: HECN NetGroup
Date: Thu, 16 Sep 2004 18:38:46 -0400 (EDT)
Cc: John.E.Dorband, Udaya.A.Ranawake,
    Josephine Palencia, Glen Gardner

[resend because a number of e-mail addresses were wrong in my first attempt to send this out]

This is a note to document the initial 10-GigE network performance testing of the thunderhead2 cluster using nuttcp. The configuration for the testing was as follows:

```
        ethn228 +-----+            +-----+            +-----+ ethn245
        ethn229 |     |            |     |            |     | ethn246
        ethn230 |     |            |     |            |     | ethn248
        ethn231 |     |            |     |            |     | ethn249
        ethn233 |     |            |     |            |     | ethn250
        ethn234 |     |            |     |            |     | ethn251
        ethn235 |  E  |            |  E  |            |  E  | ethn252
        ethn236 |  3  +============+  3  +============+  3  | ethn253
        ethn237 |  0  |  10-GigE   |  0  |  10-GigE   |  0  | ethn254
        ethn238 |  0  |            |  0  |            |  0  | ethn255
        ethn239 |     |            |     |            |     | ethn256
        ethn240 |     |            |     |            |     | ethn257
        ethn241 |     |            |     |            |     | ethn258
        ethn242 |     |            |     |            |     | ethn260
        ethn243 |     |            |     |            |     | ethn261
        ethn244 +-----+            +-----+            +-----+ ethn262
```

The left and right Force10 E300 GigE switches are in actuality 2 separate VLANs on the same physical E300 switch (located in 28/S214), while the middle E300 switch (located in 28/W220) is used to interconnect the 2 VLANs via a 10-GigE network path.

All the nodes are:

> Dual Intel(R) Xeon(TM) CPU 2.40GHz with 1 GB memory
> running Linux kernel 2.4.22-1.2188.nptlsmp (Red Hat Linux 3.2.3-6),
> using Intel Corp. PRO/1000 XT Server Adapters
> (but not currently set for 9000 MTU jumbo frames)

All tests were for 60 seconds using a 2 MB TCP window.

Thunder2 Cluster Testing with nuttcp via 10-GigE (within Bldg 28):

No Jumbo Frames:

GigE 10 streams:

```
tx228-245:  6735.4066 MB /  60.00 sec =  941.6473 Mbps 30 %TX 80 %RX
tx229-246:  6735.2853 MB /  60.00 sec =  941.6536 Mbps 30 %TX 80 %RX
tx230-248:  6735.4111 MB /  60.00 sec =  941.6688 Mbps 33 %TX 88 %RX
tx231-249:  6735.4346 MB /  60.00 sec =  941.6793 Mbps 31 %TX 77 %RX
tx233-250:  6735.6350 MB /  60.00 sec =  941.6815 Mbps 32 %TX 86 %RX
tx234-251:  6734.9643 MB /  60.00 sec =  941.6455 Mbps 37 %TX 80 %RX
tx235-252:  6735.3750 MB /  60.00 sec =  941.6654 Mbps 30 %TX 80 %RX
tx236-253:  6735.5070 MB /  60.00 sec =  941.6670 Mbps 32 %TX 88 %RX
tx237-254:  6735.7742 MB /  60.00 sec =  941.7170 Mbps 31 %TX 88 %RX
tx238-255:  6735.6017 MB /  60.00 sec =  941.6849 Mbps 38 %TX 77
```

```
%RX

Total TX 9416.71 Mbps

rx228-245:  6735.4855 MB /  60.00 sec =  941.6932 Mbps 35 %TX 88
%RX
rx229-246:  6735.4497 MB /  60.00 sec =  941.6677 Mbps 32 %TX 87
%RX
rx230-248:  6735.4495 MB /  60.00 sec =  941.6727 Mbps 29 %TX 88
%RX
rx231-249:  6735.4645 MB /  60.00 sec =  941.6656 Mbps 30 %TX 88
%RX
rx233-250:  6735.4538 MB /  60.00 sec =  941.6895 Mbps 29 %TX 71
%RX
rx234-251:  6735.3584 MB /  60.00 sec =  941.6167 Mbps 33 %TX 88
%RX
rx235-252:  6735.6687 MB /  60.00 sec =  941.6938 Mbps 31 %TX 80
%RX
rx236-253:  6735.4708 MB /  60.00 sec =  941.6786 Mbps 25 %TX 88
%RX
rx237-254:  6735.4049 MB /  60.00 sec =  941.6692 Mbps 21 %TX 88
%RX
rx238-255:  6735.4943 MB /  60.00 sec =  941.6814 Mbps 31 %TX 88
%RX

Total RX 9416.73 Mbps

tx228-245:  4706.9375 MB /  60.00 sec =  658.0504 Mbps 29 %TX 96
%RX
rx228-245:  5458.3079 MB /  60.00 sec =  763.1328 Mbps 35 %TX 94
%RX
tx229-246:  4752.0599 MB /  60.00 sec =  664.3835 Mbps 31 %TX 94
%RX
rx229-246:  5430.9375 MB /  60.00 sec =  759.2812 Mbps 32 %TX 95
%RX
tx230-248:  4945.0625 MB /  60.00 sec =  691.3645 Mbps 29 %TX 96
%RX
rx230-248:  5143.6873 MB /  60.00 sec =  719.1320 Mbps 32 %TX 95
%RX
tx231-249:  4927.0000 MB /  60.00 sec =  688.8437 Mbps 31 %TX 97
```

```
%RX
rx231-249:  5105.3410 MB /  60.00 sec =  713.7571 Mbps 30 %TX 96
%RX
tx233-250:  4813.0996 MB /  60.00 sec =  672.9057 Mbps 27 %TX 78
%RX
rx233-250:  4366.1875 MB /  60.00 sec =  610.4384 Mbps 24 %TX 86
%RX
tx234-251:  4761.3125 MB /  60.00 sec =  665.7041 Mbps 27 %TX 96
%RX
rx234-251:  5466.3125 MB /  60.00 sec =  764.1955 Mbps 33 %TX 94
%RX
tx235-252:  4989.1875 MB /  60.00 sec =  697.5265 Mbps 29 %TX 97
%RX
rx235-252:  5081.3476 MB /  60.00 sec =  710.4223 Mbps 31 %TX 96
%RX
tx236-253:  4901.5990 MB /  60.00 sec =  685.2666 Mbps 29 %TX 96
%RX
rx236-253:  5225.1777 MB /  60.00 sec =  730.5342 Mbps 32 %TX 96
%RX
tx237-254:  4901.5625 MB /  60.00 sec =  685.2763 Mbps 30 %TX 96
%RX
rx237-254:  5053.9133 MB /  60.00 sec =  706.5719 Mbps 32 %TX 96
%RX
tx238-255:  4901.7500 MB /  60.00 sec =  685.2864 Mbps 30 %TX 96
%RX
rx238-255:  5226.2064 MB /  60.00 sec =  730.6738 Mbps 31 %TX 96
%RX
```

```
Total TX 6794.61 Mbps Total RX 7208.14 Mbps Grand Total 14002.75
Mbps
```

GigE 16 streams:

```
tx228-245:  4201.8688 MB /  60.00 sec =  587.4359 Mbps 15 %TX 48
%RX
tx229-246:  4212.8595 MB /  60.00 sec =  588.9894 Mbps 13 %TX 48
%RX
tx230-248:  4211.1250 MB /  60.00 sec =  588.7432 Mbps 15 %TX 51
%RX
tx231-249:  4227.1327 MB /  60.00 sec =  590.9927 Mbps 13 %TX 47
```

```
%RX
tx233-250:  4203.1240 MB /  60.00 sec =  587.6102 Mbps 13 %TX 49
%RX
tx234-251:  4215.7484 MB /  60.00 sec =  589.4182 Mbps 14 %TX 49
%RX
tx235-252:  4200.2448 MB /  60.00 sec =  587.2247 Mbps 15 %TX 46
%RX
tx236-253:  4225.2408 MB /  60.00 sec =  590.7206 Mbps 13 %TX 49
%RX
tx237-254:  4220.5499 MB /  60.00 sec =  590.0663 Mbps 15 %TX 49
%RX
tx238-255:  4218.2796 MB /  60.00 sec =  589.7464 Mbps 14 %TX 47
%RX
tx239-256:  4210.4375 MB /  60.01 sec =  588.6033 Mbps 15 %TX 47
%RX
tx240-257:  4229.1157 MB /  60.00 sec =  591.2708 Mbps 15 %TX 48
%RX
tx241-258:  4207.9849 MB /  60.00 sec =  588.3126 Mbps 14 %TX 47
%RX
tx242-260:  4202.9017 MB /  60.00 sec =  587.6392 Mbps 14 %TX 49
%RX
tx243-261:  4199.3306 MB /  60.00 sec =  587.1072 Mbps 13 %TX 46
%RX
tx244-262:  4192.9591 MB /  60.00 sec =  586.2171 Mbps 15 %TX 47
%RX

Total TX 9420.10 Mbps

rx228-245:  4202.7500 MB /  60.00 sec =  587.5796 Mbps 15 %TX 47
%RX
rx229-246:  4204.6707 MB /  60.00 sec =  587.8434 Mbps 15 %TX 48
%RX
rx230-248:  4219.3775 MB /  60.00 sec =  589.8986 Mbps 13 %TX 48
%RX
rx231-249:  4215.1739 MB /  60.00 sec =  589.3056 Mbps 13 %TX 46
%RX
rx233-250:  4177.6115 MB /  60.00 sec =  584.0694 Mbps 11 %TX 47
%RX
rx234-251:  4206.3250 MB /  60.00 sec =  588.0522 Mbps 12 %TX 48
%RX
```

```
rx235-252:  4202.0000 MB /  60.00 sec =  587.4724 Mbps 15 %TX 48
%RX
rx236-253:  4216.4402 MB /  60.00 sec =  589.4988 Mbps 12 %TX 47
%RX
rx237-254:  4207.3179 MB /  60.00 sec =  588.2181 Mbps 15 %TX 48
%RX
rx238-255:  4209.3755 MB /  60.00 sec =  588.5043 Mbps 14 %TX 47
%RX
rx239-256:  4225.3555 MB /  60.00 sec =  590.7880 Mbps 14 %TX 48
%RX
rx240-257:  4216.9829 MB /  60.00 sec =  589.5779 Mbps 13 %TX 47
%RX
rx241-258:  4225.1815 MB /  60.00 sec =  590.7215 Mbps 14 %TX 50
%RX
rx242-260:  4218.9673 MB /  60.00 sec =  589.8253 Mbps 13 %TX 48
%RX
rx243-261:  4229.6626 MB /  60.00 sec =  591.3385 Mbps 12 %TX 49
%RX
rx244-262:  4199.6966 MB /  60.00 sec =  587.1729 Mbps 15 %TX 47
%RX

Total RX 9419.87 Mbps

tx228-245:  4136.8978 MB /  60.00 sec =  578.3744 Mbps 21 %TX 67
%RX
rx228-245:  4146.9375 MB /  60.00 sec =  579.7778 Mbps 22 %TX 66
%RX
tx229-246:  4131.6233 MB /  60.00 sec =  577.6394 Mbps 22 %TX 68
%RX
rx229-246:  4141.3291 MB /  60.00 sec =  578.9879 Mbps 22 %TX 65
%RX
tx230-248:  4104.4361 MB /  60.00 sec =  573.8293 Mbps 23 %TX 67
%RX
rx230-248:  4120.1189 MB /  60.00 sec =  576.0208 Mbps 22 %TX 66
%RX
tx231-249:  4129.1645 MB /  60.00 sec =  577.2842 Mbps 22 %TX 68
%RX
rx231-249:  4122.5000 MB /  60.00 sec =  576.3604 Mbps 21 %TX 67
%RX
tx233-250:  4017.4267 MB /  60.00 sec =  561.6696 Mbps 19 %TX 61
```

```
%RX
rx233-250:  3857.8125 MB /  60.00 sec =  539.3572 Mbps 18 %TX 68
%RX
tx234-251:  4137.9375 MB /  60.00 sec =  578.5336 Mbps 22 %TX 68
%RX
rx234-251:  4144.4854 MB /  60.00 sec =  579.4146 Mbps 21 %TX 66
%RX
tx235-252:  4135.2286 MB /  60.00 sec =  578.1462 Mbps 21 %TX 67
%RX
rx235-252:  4143.2553 MB /  60.00 sec =  579.2589 Mbps 22 %TX 67
%RX
tx236-253:  4122.9369 MB /  60.00 sec =  576.4209 Mbps 21 %TX 68
%RX
rx236-253:  4124.5625 MB /  60.00 sec =  576.6502 Mbps 22 %TX 67
%RX
tx237-254:  4121.8659 MB /  60.00 sec =  576.2668 Mbps 21 %TX 67
%RX
rx237-254:  4123.6763 MB /  60.00 sec =  576.5194 Mbps 21 %TX 67
%RX
tx238-255:  4145.1113 MB /  60.00 sec =  579.5257 Mbps 22 %TX 67
%RX
rx238-255:  4149.5192 MB /  60.00 sec =  580.1342 Mbps 22 %TX 68
%RX
tx239-256:  4152.4375 MB /  60.01 sec =  580.4954 Mbps 21 %TX 68
%RX
rx239-256:  4163.6189 MB /  60.00 sec =  582.1520 Mbps 21 %TX 67
%RX
tx240-257:  4134.4369 MB /  60.00 sec =  578.0229 Mbps 21 %TX 66
%RX
rx240-257:  4124.8125 MB /  60.00 sec =  576.6953 Mbps 21 %TX 68
%RX
tx241-258:  4129.6668 MB /  60.00 sec =  577.3690 Mbps 21 %TX 69
%RX
rx241-258:  4139.1250 MB /  60.00 sec =  578.6796 Mbps 21 %TX 67
%RX
tx242-260:  4109.8108 MB /  60.00 sec =  574.6119 Mbps 21 %TX 67
%RX
rx242-260:  4126.8159 MB /  60.00 sec =  576.9315 Mbps 21 %TX 67
%RX
tx243-261:  4117.9375 MB /  60.00 sec =  575.7275 Mbps 21 %TX 67
```

```
%RX
rx243-261:  4128.5625 MB /   60.00 sec =  577.1980 Mbps 21 %TX 66
%RX
tx244-262:  4115.8687 MB /   60.00 sec =  575.4344 Mbps 21 %TX 68
%RX
rx244-262:  4120.8125 MB /   60.00 sec =  576.1252 Mbps 21 %TX 67
%RX

Total TX 9219.35 Mbps Total RX 9210.26 Mbps Grand Total 18429.61
Mbps
```

That's all for now.

-Bill